EHRA
HiMSS Electronic Health Record Association

Balancing the Need for Privacy and the Value of Big Data
EHR Association Response to the PCAST Report, *Big Data and Privacy: A Technological Perspective*

The President's Council of Advisors on Science and Technology (PCAST) published a report in May 2014 on the topic of ***Big Data and Privacy: A Technological Perspective***. Representing the nearly 40 members of the Electronic Health Record Association (EHRA), we appreciate the publication of this PCAST report as it addresses an increasingly important question as data is being aggregated in various forms that have the risk of compromising an individual's privacy while providing potential benefits to society as a whole. This issue is of particular importance and interest to the healthcare industry, where great care is given to protect patient data and where re-identification of patient data in secondary data use is of concern. This interest is evidenced by Health Insurance Portability and Accountability Act (HIPAA) and its updates, which established key criteria and expectations on how to protect patient data that must be taken into consideration when addressing big data and privacy.

The EHRA conducted a review of the PCAST report and makes the following observations, in the context of "big" health data. Our reviewers considered the increased aggregation of health data to enable a variety of analyses and research to derive new knowledge from the aggregated data. We see this report as relevant to discussions and direction on how to best gain public health value from these vast data sources, both health data-specific and more general socio-economic data, while maintaining patient privacy and confidentiality. This topic is of further interest as the recently released JASON report[1] also addresses the needs and opportunities to weave big data into the healthcare information fabric. Therefore, the following observations offer a starting point for further discussions among EHRA and its members, providers, policymakers, and other stakeholders to find a practical balance of guidance (statutory, regulatory, and/or standards) to maintain essential patient privacy while enabling the benefits of big data in healthcare.

Overall, the report establishes an important and clear perspective on the scope and reach of the uses of health data in the "big data" space as it exists today and as it may evolve in the future. We agree that it accurately frames the extent of challenges that the industry and policymakers should consider (i.e., the implications of big data in terms of individual privacy concerns vs. the potential value that can be derived from big data, and how to approach privacy protection within technology capabilities). In addition, given the speed of technological advances, we agree that the proposed privacy protection policy approach should not focus on technical solutions alone but rather be stated in terms of intended outcomes. Our specific thoughts are outlined below.

[1] A Robust Health Data Infrastructure, Prepared for the Agency for Healthcare Research and Quality by JASON, The MITRE Corporation, AHRQ Publication No. 14-0041-EF, April 2014

Protection of Patient Data

This PCAST report states in its cover letter that big data drives big benefits, from innovative businesses to new ways to treat diseases. The EHRA recognizes and supports this notion as our members, to a greater or lesser extent, have enabled providers to tap into the benefits of data analytics. Health data can inform the analytical activities that occur within data registries and research settings. In many cases, uses of health data can add exceptional value in the learning healthcare systems that are being designed and built today. Generally, such uses of patient information are not based on the potential benefits for any given individual (e.g., clinical decision support (CDS)-driven personalized medicine), but rather are focused on improving the health and healthcare for larger populations and establishing the knowledge that enables CDS-driven personalized medicine. The needed balance is between protecting the rights of the individual and improving the healthcare of the larger US population. Where these kinds of activities do not use individually identifiable health data, we believe that they should not be arbitrarily restricted or subject to requirements to gain individual consent. Since the adoption of HIPAA in 1996, significant focus has been placed on the protection of individual patient data in healthcare IT, as well as sharing the minimum necessary data set for the intended use. While both technology and perspectives on privacy have advanced or changed since that time, a clear approach was established on how to protect patient identity within healthcare.

As a general matter, with certain exceptions, no use of identifiable health data (i.e., data that enables the user of that data to establish the identity of the patient who is the subject of that data) is allowable without patient consent or legal authorization consistent with the HIPAA regulations. Covered entities – e.g., health plans, healthcare clearinghouses, and healthcare providers who electronically transmit any health data in connection with transactions for which the Department of Health and Human Services **(**HHS) has adopted standards – are subject to HIPAA for stewardship over individually identifiable health data.

The associated regulations are complex and, among other aspects, address the practical aspects of treatment, payment, and healthcare operations. Despite these complexities and challenges, any solution to protect patient privacy moving forward must consider the lessons learned from the HIPAA regulation as well as account for the need for the patient to appropriately consent to the use of data about them for purposes beyond their direct care.

Unfortunately, neither this PCAST report nor the JASON report addresses this existing HIPAA framework in any real detail, nor acknowledges clearly that technology alone is not sufficient to achieve the desired outcome of protecting individual patient privacy while enabling big data benefits. Appropriate policy and individual/organizational behaviors are essential as well. We believe it is important to understand the level of control that patients have over their data, the right of a patient to have a copy of data about them, the level of control the provider has over their patients' data, as well as the rules that apply to the original data vs. copies of such data to determine how to appropriately apply policy, technology, and behaviors. We recognize that technology has evolved and perspectives on privacy have changed in general, but we suggest it is important to address these questions specifically in the context of health data and existing laws before establishing any new regulatory controls and applying new technologies.

- **Born Digital and Born Analog**
  This PCAST report establishes a categorization of data as either born-digital or born-analog. Data born-digital is defined as data that is specifically collected for use by a computer, while data born-analog is defined as data when collected through sensors such as a camera, microphone, or similar devices. While this is an interesting distinction, the concept is not carried forward in a way that can help address potential challenges. We suggest that, particularly in the health data space, data that is "born digital" is exploding with the advent of personal devices and other advances in diagnostic/sensing devices (likely accelerating the need for "big data" tools and techniques), while data born analog continues to rapidly expand as well. But, on balance, as data are aggregated and combined, its origins should not affect the privacy issues being addressed. Consequently, we suggest that distinguishing data along these axes is not helpful in addressing the fundamental privacy concerns involving health data.

- **Patient Re-Identification**
  This PCAST report identifies anonymization as somewhat useful, but considers it is not robust against near-term future re-identification methods. The EHRA agrees that, increasingly, data can be re-constructed into identifiable data as long as there is enough of it to establish a pattern – whether the data itself or its metadata is with or without protected health information (PHI) in the initial data set. As a result, even deletion of data while retaining metadata may become problematic from a privacy perspective. Consequently, the ability of an individual to control access to and use of their identifiable (directly or indirectly) health data remains of great importance. While the report does address the issue of re-identifying data, specifically in the context of healthcare data, we suggest that it requires more focus in the context of healthcare to ensure the overall complexities are addressed (e.g., given sufficient volume, even metadata can be used to re-identify individuals).

- **Anonymization vs. De-Identification**
  The PCAST report obscures the distinction between anonymization and de-identification, whereas standards such as ISO 25237 have specific definitions for these terms. We suggest that further discussions are needed to clarify the definitions and use established sources like HIPAA and the new ISO definitions.

  The report describes challenges associated with anonymization, claiming that new analysis techniques increase the risk of re-identification. We note that de-identification does not promise perfect privacy—it simply establishes that the risk of re-identification is low. As analysis techniques become more advanced, the risk of re-identification changes and de-identification techniques need to be updated to keep pace. This reality does not mean that de-identification is an ineffective privacy control.

The report cites re-identification work on the Personal Genome Project by Sweeney et al. (http://dataprivacylab.org/projects/pgp/index.html), which included zip code and birthdate as data elements. The HIPAA Privacy Rule has clear processes for creating a de-identified data set and this data set would not be considered as “de-identified” according to the HIPAA Privacy Rule. Despite the theoretical risks, we are unaware of any published work that demonstrates that data that is de-identified according to the HIPAA-standard can be defeated. De-identified data is of key importance in performing clinical trials research, population health management, and syndromic and public health surveillance. De-identification is a key privacy control in these contexts that allows personal health data to be used for broader societal goals (following patient consent preferences), but balance is not clear in the report. However, considering there is still a potential to re-identify individuals and the challenges to keep up with appropriate de-identification mechanisms, it is essential that organizations managing de-identified health data establish clear, verifiable, and perhaps certified processes to improve on managing the risks of inappropriate re-identification.

- **From Notice and Consent to Use Control**
  The PCAST report indicates that the purpose of notice and consent is that the user assents to the collection and use of personal data for a stated use that is acceptable to that individual. It then asserts that, given the large number of programs and Internet-available devices, both visible and not, that collect and use personal data, this framework is increasingly unworkable and ineffective. The report argues that that all data identified as potentially needed must be contributed to data-aggregators (e.g., researchers, registries, etc.) without restriction and with (limited) controls, thus moving from a notice-and-consent model to a use-control model. We do not believe this principle is justified in all or even most situations, particularly in healthcare.

  Overall, the report oversimplifies the ability to only focus on use controls. Individuals should continue to be able to restrict contributing their personally identifiable health data (or other data consistent with organizational policies) to the big data pool; or, where such data has value at a population level, great care must be taken to protect the contributing individuals’ identities. Until an individual can be assured that their data will be kept secure and private, they should not be required to contribute. This issue is of particular concern when and if data is uploaded that should not have been and the disclosure event is not reversible, such as behavioral health data having accidently been shared through a health information exchange (HIE), or forwarded from one provider to another, accidently copying another. The individual’s privacy has essentially been compromised forever.

  The report also suggests minimizing potential harm through addressing appropriate use of data. Unfortunately, that approach may not substantially reduce the challenges to maintain patient privacy. There is still a need for consent management to establish the purposes for which data can or cannot be used above and beyond treatment, payment, and healthcare operations (TPO). Once the data is available, the risk may increase despite having use controls in place; it may still be used incorrectly, thus creating irreversible harm. It is important to recognize that the most serious type of privacy breach is the accidental release of patient data.

Without rigorous protections in place to prevent those types of breaches, we should not move too quickly to an exclusively use-based regulatory framework. Given the increasing availability of data, revisiting the feasibility of a use-based model in the future will be worthwhile. Regardless of the model, the regulations must not be so restrictive that they become an obstacle to data sharing and use for patient treatment.

We believe that this report should have addressed more clearly the many concerns with individuals' contribution of their identifiable health data, particularly in the healthcare space. In addition, the report could have included a clearer overall discussion of the challenges and current regulations in the healthcare industry that preclude or limit the use of big data. Additionally, it would be helpful to further address the potential penalties of inappropriate use/disclosure to clarify the severity of such use/disclosures, particularly when it involves health data.

- **User Preferences**
  The PCAST report authors believe that the responsibility for using personal data in accordance with the user's preferences should rest with the provider, possibly assisted by a mutually accepted intermediary, rather than with the user. Although we agree with the concept of an intermediary or "data manager" (provider, HIE, federal or state agency, etc.) that would have clear responsibility to use an individual's data in accordance with their preferences, we suggest that including the preference not to share certain data with other, aggregated data managers, is important as well, particularly when such data can identify the individual. This role is clearly addressed by HIPAA.

  We suggest that it would be helpful to provide examples of use-control models to improve understanding of what should be accomplished by these models and demonstrate the effectiveness of such models. We suggest that clarifying data provenance via metadata would empower the source to provide use-control criteria so subsequent users can then manage the data appropriately. We do suggest that for health data, such metadata is kept at a higher level of aggregation (e.g., a clinical summary document). This approach can avoid the many complexities of such controls at more granular data element levels, particularly the need for more complex administrative controls and the need to review with the individual the implications of their choices. To provide consent, the individual must be well informed. Proper notices and transparency on intended use are essential to enable an individual to make a consent decision that is right for them. It is unclear whether the benefits of controls at more granular data element levels outweigh the challenges to properly manage the consent process.

Furthermore, there should be controls across all three roles mentioned in the report – data collectors, data analyzers, and users of the analyzed data – to ensure that certain data is not collected and inadvertently released. Clear definitions will be needed for what data should be made available to which data analyzers. Such a form of consent could still be based on general models, rather than the currently dominant take-it-or-leave it approach.

- **Privacy Preference Profiles**
  The report suggests the use of third parties to manage use criteria through privacy preference profiles. We suggest that such a model will not work well in healthcare. The need for providers and patients to communicate with other providers affiliated with third parties who are managing use criteria will be unnecessarily complicated. This complexity is already demonstrated with the different approaches between states on managing privacy and security, resulting in substantial challenges when data needs to cross state lines. Rather, one model should be agreed upon so that providers and patients can consistently exchange and aggregate data under the same guidelines.

- **A Roadmap Towards a Use Control Model**
  This PCAST report does not provide a roadmap on how to switch from a "control of the collection" paradigm to a "control of the use" paradigm. We suggest that this shift will take substantial time, trust, and investment. Before this new approach can be successfully applied to health data, we suggest that trust in the proposed method must first be established and validated in the general, commercial data space. We note that even in the commercial and social media space, unexpected use of data raises substantial trust issues. We agree that the notion of focusing on abuse of use is an important component of addressing the problem statement at large but suggest that, until sufficient trust is established first in such controls, we cannot solely rely on that approach within healthcare for some time. Therefore, we propose extension of the first recommendation that currently states "*Policy attention should focus more on the actual uses of big data and less on its collection and analysis",* to include *"In healthcare, collection should remain focused on the minimum data that is necessary for the intended use with appropriate patient consent."*

General Comments
Based on its analysis, this PCAST report recommends further research on privacy-related technologies. Such research will be necessary to maintain the privacy of health data, while enabling the larger population benefits that big data research and analysis can offer. This additional research should address management of provenance, consent, and data segmentation. It is important to identify minimum data sets for a variety of intended uses – population health management, research, and outcomes analyses. It is equally important to arrive at a practical level of data granularity to which controls will be applied. We suggest focus on document-level granularity to establish capabilities and experience before considering lower levels of granular controls.

Throughout this research and development of new policies, it is essential to maintain a practical balance between the needs to share data for delivery of patient care, secondary/research data use, individual privacy protection, and public health objectives. Bringing together all stakeholders will enable us to maintain such balance as technology and the perspectives on privacy evolve, and individuals' participation in their healthcare increases.

Finally, although this report provides a vision of shifting from focus on data collection to focus on data use, we anticipate that in healthcare both areas of focus will be continue to be important and interact in complex ways. Until privacy concerns have been appropriately addressed to limit the chance of irreversible harm through disclosures, a fully use-based model will not emerge. We look forward to being a part of the important collaborative effort to resolve these issues.

About the EHR Association
Established in 2004, the Electronic Health Record (EHR) Association is comprised of nearly 40 companies that supply the vast majority of operational EHRs to physicians' practices and hospitals across the United States. EHRA operates on the premise that the rapid, widespread adoption of EHRs will help improve the quality of patient care as well as the productivity and sustainability of the healthcare system as a key enabler of healthcare transformation. The EHR Association and its members are committed to supporting safe healthcare delivery, fostering continued innovation, and operating with high integrity in the market for our users and their patients and families.

The EHR Association is a partner of the Health Information Systems Management Society (HIMSS). For more information, visit www.ehrassociation.org.